JParaCrawl v3.0: 大規模日英対訳コーパス

森下 睦 [1,2], 帖佐 克己 [1], 鈴木 潤 [2], 永田 昌明 [1]
[1]NTT コミュニケーション科学基礎研究所　[2] 東北大学
makoto.morishita.gr@hco.ntt.co.jp

概要

現在の機械翻訳モデルは主に対訳コーパスを用いて学習されており、その翻訳精度は対訳コーパスの質と量に大きく依存している。本稿では、新たにウェブをクロールし日英対訳文を抽出することで大規模日英対訳コーパスを構築し、翻訳精度の底上げを狙う。今回ウェブから収集した対訳文と以前作成した日英対訳コーパス JParaCrawl v2.0 を合わせることで、合計 2100 万文を超える日英最大規模の対訳コーパスを作成することに成功した。実験により、新たな対訳コーパスを用いて学習した翻訳モデルが様々な分野で高い翻訳精度を発揮することを示す。なお、今回作成した対訳コーパスを JParaCrawl v3.0 と名付け、我々のウェブサイト上で研究目的利用に限り無償公開する予定である。

1　はじめに

現在のニューラル機械翻訳モデルは、主に対訳コーパスを用いた教師ありの手法 [1, 2, 3, 4] で学習されている。学習時の対訳コーパスの質と量が翻訳精度に大きな影響を与えること知られているが、一般に公開されている対訳コーパスは多くの言語対で限られている。例えば、独英などの言語対ではすでに数億文の対訳文が公開されているものの、日英ではまだ同程度のものは存在せず、翻訳モデル学習時に大きな問題となっている。そのため、本稿ではさらに大規模なウェブベースの日英対訳コーパスを構築する。現在、日英で最大規模の対訳コーパスの 1 つは約 1000 万文の対訳文を含む JParaCrawl v2.0 [5] であり、これはウェブを大規模にクロールし対訳文を自動的に抽出することで構築されている。本コーパスは欧州言語対と比較すると小規模であり、2020 年を最後に更新が止まっているため、最新の情報を含んでいない。そのため、本研究ではウェブを全面的に再クロールし、対訳文を抽出することで JParaCrawl コーパスを拡大/更新する。本研究では、対訳文抽出手法を機械翻訳器を用いた手法へと変更し、新たに PDF や Word 文書も収集対象とすることで、対訳抽出数の向上を狙う。また、新たに作成した対訳コーパスを用いて、英日および日英の機械翻訳の精度がどのように向上するかを実験的に示す。本研究で作成した対訳コーパスは JParaCrawl v3.0 と名付け、今後の研究のために我々のウェブサイト[1)] で公開する予定である。

本稿の貢献は以下のようにまとめられる。

- 従来の JParaCrawl コーパスと合わせて 2100 万文対以上を含む大規模な日英対訳コーパスを構築した。
- 新たな対訳コーパスにより、幅広い分野で英日・日英機械翻訳の精度が向上することを実験的に確認した。
- 本コーパスを研究目的利用に限り無償で公開する。

2　関連研究

対訳コーパスは様々な文書から対訳文を抽出することで作成されることが多い。代表的なものとして、国際機関が作成した対訳文書がある。例えば、欧州議会の議事録から作成された Europarl [6]、国連の翻訳文書から作成された UN 対訳コーパス [7] などがある。これらの文書は、通常プロの翻訳者が翻訳しており、文書 ID などのメタデータを持っていることもあるので、容易に対訳文を抽出することができる。しかし、これらの一般に公開されている対訳文書は多くない。

近年では、ウェブから対訳文を抽出する手法も多く提案されている。ウェブ上には 2 言語以上で書かれたウェブサイトが多数存在し、こういったウェブサイトから対訳文を抽出する。ウェブ上には、様々な言語や分野の対訳文が存在しており、大規模な

1) http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

表 1 JParaCrawl コーパスに含まれる重複を取り除いた対訳文数および英語側単語数

バージョン	文数	単語数
v1.0	4,817,172	125,216,523
v2.0	8,809,771	234,393,978
v3.0	21,481,513	502,445,763

対訳コーパスを作成するための有望な情報源である。ウェブから対訳文を抽出する初期の研究としては、大規模な分散システムを構築し対訳文を抽出したもの [8]、Common Crawl[2] から対訳文を抽出したもの [9] などがある。最近では、多言語文埋め込みを用いた対訳文対応手法を用いて、Wikipedia や Common Crawl から大規模な多言語対訳コーパスを作成する研究も報告されている [10, 11]。

また、ParaCrawl プロジェクトはヨーロッパ言語の大規模な対訳コーパスをウェブから継続的に作成している [12]。我々は以前この活動にヒントを得て、大規模な対訳コーパスが存在しない日英向けの大規模な対訳コーパスを作成した [5]。このコーパスは JParaCrawl と名付けられ、1,000 万文を超える対訳文を含む日英における最大規模の対訳コーパスとなっている。しかし、JParaCrawl コーパスは、独英などの欧州言語対と比較するとまだ小規模であり、これをもとにした翻訳モデルの精度も欧州言語対と比較すると低精度である。ゆえに、さらに大きな日英対訳コーパスの作成が求められている。本研究では、ウェブを新たにクロールし対訳文をさらに抽出することで、JParaCrawl コーパスをさらに拡張することを目指す。

3 JParaCrawl v3.0

本研究では、さらにウェブをクロールして対訳文を抽出し、JParaCrawl v2.0 コーパスを拡張すること目指す。我々の手法は、以前の ParaCrawl および JParaCrawl プロジェクトに基づくものである。以下の節でその詳細な手順を述べる。

3.1 対訳文を含むウェブサイトの発見

本研究では、ウェブから対訳文を抽出することで大規模な対訳コーパスを構築する。まず、CommonCrawl 上のテキストデータを `CLD2`[3] によって解析し、各ドメインの言語別データ量を得る。その後、英語と日本語が同量程度含まれるウェブサイトには対訳文が存在する可能性があるという仮説に基づき、クロール対象ウェブサイトを列挙する。本研究では、2019 年 3 月から 2021 年 8 月までに公開された Common Crawl のテキストアーカイブデータを分析対象とし、ウェブサイトの規模が大きく、英語と日本語の文章が同程度である 10 万件のウェブサイトを列挙した。2019 年 3 月以前に公開されたデータについては、JParaCrawl v2.0 作成時に既に分析済みであり、本収集の目標の一つである最新の情報を含むこととも乖離しているため除外した。本手順で列挙されたウェブサイト一覧を確認したところ、前回の JParaCrawl v2.0 時に候補に挙がっていないウェブサイトが 7 割を占めていた。なお、本手順には ParaCrawl プロジェクトが提供する `extractor`[4] を使用した。

3.2 ウェブサイトのクロール

次に、前節で列挙されたウェブサイト全体をクロールする。本研究では、`Heritrix`[5] を用いて、各ウェブサイトに対して最大 48 時間のクローリングを行った。これまでの JParaCrawl では、プレーンテキストのみを対象としていたが、今回はこれに加えて、PDF や Word 文書もクロールの対象とした。日本国内の官公庁や企業は PDF で文書を発信することがあり、これらも対訳抽出の対象とすることで対訳文数の増加に寄与すると考えられる。

3.3 対訳文抽出

次に、クロールされたウェブサイトから対訳文を抽出する。本手順には、ParaCrawl プロジェクトが提供する `Bitextor`[6] を日本語に対応させ使用した。対訳文書と対訳文の対応付けには、機械翻訳を用いた対応付けツール `bleualign` [13] を使用した。このツールでは、まず機械翻訳を用いて日本語文を英文に翻訳し、BLEU スコアを最大化する日英の文ペアを発見する。この際、日英翻訳には JParaCrawl v2.0 で学習した Transformer ベースのニューラル機械翻訳（NMT）モデルを使用した。なお、以前の JParaCrawl で使用した辞書ベースの手法 [14] よりも `bleualign` の方が高精度であることを予備実験により確認した。

2) https://commoncrawl.org/
3) https://github.com/CLD2Owners/cld2
4) https://github.com/paracrawl/extractor
5) https://github.com/internetarchive/heritrix3
6) https://github.com/bitextor/bitextor

3.4 ノイズ除去

最後の手順として、正しく対応付けられていない、翻訳が不正確など、学習時のノイズとなる文対をフィルタリングする。本手順には、Bicleaner[7] [15] を使用した。ノイズ除去の後、クリーンな対訳文と JParaCrawl v2.0 を結合、重複文を削除した。以上の手順により、2100 万文以上を含む新しい大規模な JParaCrawl v3.0 を作成した。これは、以前の JParaCrawl v2.0 の倍以上の文数である。

表 1 に、これまでの JParaCrawl および今回作成した v3.0 の重複を削除した対訳文数および英語側単語数を示す。なお、これまでの JParaCrawl についても重複削除を行った文数を報告しているため、以前の論文で報告されている値とは異なることに注意されたい。PDF なども対訳抽出対象としたことにより、HTML のみを対象とする場合と比較して、約 10%対訳文抽出数が向上した。

4 実験

本節では、新たに作成した JParaCrawl v3.0 の翻訳精度への影響を確認するために、NMT モデルを学習し様々なテストセットでその精度を評価した。以降では、使用したテストセットの詳細およびモデル学習時の設定について述べる。

4.1 実験設定

4.1.1 テストセット

様々な分野で NMT モデルの精度を評価するために、15 種のテストセットでモデルを評価する。付録表 3 に使用するテストセットの分野および統計情報を示す。これらには、以前の JParaCrawl 発表時に報告した ASPEC [16] (科学技術論文), JESC [17] (映画字幕), KFTT [18] (Wikipedia 記事), TED (tst2015) [19] (講演) が含まれる。さらに本実験では、会話文中心の対訳コーパスである Business Scene Dialogue コーパス (BSD) [20] や翻訳シェアードタスク [21, 22, 23, 24, 25] のテストセットを追加した。これらには、ニュース記事、SNS 上のテキスト、Wikipedia 上のコメントなどが含まれる。なおシェアードタスク用テストセットの中には、特定の翻訳方向 (英→日など) で使用することを前提としたものもあるが、参考までに英日、日英の両方向で評価した。また付録表 4 に示すように、いくつかのコーパスには学習データが付属しているものがある。比較のため、これらの学習データでも分野別モデルを学習しスコアを報告する。

4.1.2 学習設定

前処理として、対訳コーパスを sentencepiece [26] を用いてサブワード単位に分割した。この際、語彙数は 32,000 とした。翻訳モデルの学習には fairseq [27] を用い、small、base、big の 3 つの異なる大きさの Transformer [4] モデルを学習した付録表 5 に詳細な学習設定を示す。以前の JParaCrawl の報告に基づき、TED (tst2015) では small モデルを、KFTT では base モデルを、その他のテストセットでは big モデルを使用した。なお、学習に関する設定は、公平な比較のために以前の JParaCrawl の報告とほぼ同じであるが、v3.0 モデルについては、対訳コーパスの大きさが原因で 24,000 ステップでは収束せず、更新回数を変更した。評価には sacreBLEU [28] を用い、BLEU スコア [29] を報告する。なお、以前の実験との整合性を保つために、すべてのテストセットを NFKC 正規化した。

4.2 実験結果

表 2 に様々なテストセットにおける BLEU スコアを示す。JParaCrawl v3.0 で学習したモデルは、日英ではすべてのテストセットで、英日では 15 のうち 13 のテストセットで以前の JParaCrawl を上回る精度を達成した。この結果は、科学技術論文、ニュース、対話などの様々な分野において、新しい JParaCrawl が NMT モデルの精度を押し上げることを示している。特に、WMT21 ニュース翻訳タスクでは、JParaCrawl v3.0 モデルが大幅に精度を向上させることがわかった。これは、前回の JParaCrawl v2.0 が 2019 年のウェブデータをもとに作成されており、2021 年のニュースに頻出する単語が正しく翻訳できていないことが原因だと推測される。例えば、2021 年のニュース記事では、2019 年には存在しない COVID-19 に関連する単語が頻出する。こういった最新の単語を正しく翻訳できるようにするためには、対訳コーパスを常に更新し続けるなどの対応が必要だと考えられる。

なお、JParaCrawl は特定の分野には特化することを目的としておらず、JParaCrawl 単独では分野別モデルの精度には及ばない。しかし、JParaCrawl モデ

7) https://github.com/bitextor/bicleaner

表 2　分野別コーパスおよび JParaCrawl v1.0, v2.0, v3.0 で学習した翻訳モデルの BLEU スコア。JParaCrawl モデルのうち最高スコアのものを太字で示す。

	英日翻訳				日英翻訳			
		JParaCrawl				JParaCrawl		
テストセット	分野別	v1.0	v2.0	v3.0	分野別	v1.0	v2.0	v3.0
ASPEC	44.3	24.7	26.5	**26.8**	28.7	18.3	19.7	**20.8**
JESC	14.5	**6.6**	6.5	6.5	17.8	7.0	7.5	**8.4**
KFTT	31.8	17.1	**18.9**	18.1	23.4	13.7	16.2	**17.0**
TED (tst2015)	11.1	11.5	12.6	**13.1**	13.7	11.0	11.9	**12.0**
Business Scene Dialogue Corpus	—	12.4	13.5	**13.9**	—	17.4	19.6	**19.9**
WMT20 News En-Ja	—	20.7	21.9	**23.5**	—	21.3	23.3	**23.9**
WMT20 News Ja-En	—	20.1	22.8	**23.5**	—	19.2	21.0	**21.9**
WMT21 News En-Ja	—	21.1	21.8	**25.0**	—	21.9	23.1	**24.3**
WMT21 News Ja-En	—	19.6	21.5	**22.4**	—	18.1	20.7	**21.3**
WMT19 Robustness En-Ja (MTNT2019)	—	12.4	12.5	**14.4**	—	15.6	16.8	**17.3**
WMT19 Robustness Ja-En (MTNT2019)	—	11.5	12.3	**12.8**	—	16.0	17.2	**17.7**
WMT20 Robustness Set1 En-Ja	—	15.2	15.8	**18.7**	—	20.0	20.6	**21.6**
WMT20 Robustness Set2 En-Ja	—	12.7	13.0	**14.8**	—	16.4	17.4	**17.9**
WMT20 Robustness Set2 Ja-En	—	7.9	8.2	**8.6**	—	12.0	12.6	**14.0**
IWSLT21 Simultaneous Translation En-Ja Dev	—	12.5	13.3	**14.5**	—	12.9	14.3	**14.5**

入力文	院内に「濃厚接触者」はいませんが、接触者全員に PCR 検査を実施し、 女性が関係した病棟などを閉鎖して徹底的に消毒するということです。
参照訳	There are no known “‘close contacts” in the hospital, but all contacts will be subjected to PCR tests, and the wards and other areas where the women had been will be closed and thoroughly disinfected.
JParaCrawl v1.0	There is no “strong contact person” in the hospital, but a PCR test will be conducted for all the contacts, and women will close the wards and thoroughly disinfect them.
JParaCrawl v2.0	Although there is no “strong contact person” in the hospital, PCR tests will be performed on all contact persons, and the wards related to women will be closed and thoroughly disinfected.
JParaCrawl v3.0	There are no “close contacts” in the hospital, but PCR tests will be conducted for all contacts, and the wards related to women will be closed and thoroughly disinfected.

図 1　WMT21 News Ja-En テストセットの翻訳例

ルから特定分野に対して Fine-tuning を行うことで、分野別データ単独で学習したモデルを上回ることも報告されている [5]。本傾向は JParaCrawl v3.0 でも同様だと思われるが、具体的な実験については今後の課題とする。

4.3　翻訳例

図 1 に JParaCrawl v1.0、v2.0、v3.0 で学習したモデルの翻訳例を示す。この例は、WMT21 ニュース翻訳テストセットから COVID-19 に関連するものを選んだ。この入力文には、「濃厚接触者」という日本語が含まれており、これは “close contacts” と翻訳されるべきである。しかし、v1.0 および v2.0 で学習したモデルは “strong contact person” と誤訳している。一方で、v3.0 で学習したモデルでは、これを正しく “close contacts” と翻訳できている。テストセットを目視で確認したところ、この例のように COVID-19 に関連する記事で多くの改善が見られた。この結果は、前節で述べた近年頻出する用語を正しく翻訳できているという仮説を支持するものである。

5　おわりに

本研究では、これまでの大規模日英対訳コーパス JParaCrawl をさらに拡張した JParaCrawl v3.0 を作成した。本対訳コーパスは、最新の CommonCrawl アーカイブを分析して対訳文が存在すると思われるウェブサイトを発見し、それらから対訳文抽出を行うことで作成した。新たな JParaCrawl v3.0 は 2100 万以上の対訳文を含んでおり、これは JParaCrawl v2.0 の倍以上の大きさである。本対訳コーパスを用いることで、様々な分野、特に最新のニュース記事の翻訳精度が向上することを実験的に確認した。今後の課題としては、継続的な JParaCrawl の更新や、より優れた対訳文抽出/フィルタリング手法の提案などが挙げられる。なお、本研究で作成した JParaCrawl v3.0 は我々のウェブサイトで研究目的に限り無償で公開する予定である。

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 3104–3112, 2014.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[3] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1412–1421, 2015.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 6000–6010, 2017.

[5] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pp. 3603–3609, 2020.

[6] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In Proceedings of the Machine Translation Summit X, pp. 79–86, 2005.

[7] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1.0. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pp. 3530–3534, 2016.

[8] Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pp. 1101–1109, 2010.

[9] Jason R Smith, Herve Saint-Amand, M Plamada, P Koehn, C Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1374–1383, 2013.

[10] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. arXiv preprint arXiv:1907.05791, 2019.

[11] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. CCMatrix: Mining billions of high-quality parallel sentences on the web. arXiv preprint arXiv:1911.04944, 2019.

[12] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4555–4567, 2020.

[13] Rico Sennrich and Martin Volk. Iterative, MT-based sentence alignment of parallel texts. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA), pp. 175–182, 2011.

[14] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing (RANLP), pp. 590–596, 2005.

[15] Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In Proceedings of the 3rd Conference on Machine Translation (WMT), pp. 955–962, 2018.

[16] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), 2016.

[17] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. JESC: Japanese-English Subtitle Corpus. arXiv preprint arXiv:1710.10639, 2017.

[18] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[19] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: web inventory of transcribed and translated talks. In Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), pp. 261–268, 2012.

[20] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In Proceedings of the 6th Workshop on Asian Translation (WAT), pp. 54–61, 2019.

[21] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In Proceedings of the 5th Conference on Machine Translation (WMT), pp. 1–55, 2020.

[22] Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In Proceedings of the 6th Conference on Machine Translation (WMT), pp. 1–93, 2021.

[23] Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir K. Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan M. Pino, and Hassan Sajjad. Findings of the first shared task on machine translation robustness. In Proceedings of the 4th Conference on Machine Translation (WMT), 2019.

[24] Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. Findings of the WMT 2020 shared task on machine translation robustness. In Proceedings of the 5th Conference on Machine Translation (WMT), pp. 76–91, 2020.

[25] Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT), pp. 1–29, 2021.

[26] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 66–71, 2018.

[27] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), pp. 48–53, 2019.

[28] Matt Post. A call for clarity in reporting BLEU scores. In Proceedings of the 3rd Conference on Machine Translation (WMT), pp. 186–191, 2018.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318, 2002.

[30] Graham Neubig. Forest-to-string SMT for asian language translation: NAIST at WAT2014. In Proceedings of the 1st Workshop on Asian Translation (WAT), pp. 20–25, 2014.

[31] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, Vol. 15, pp. 1929–1958, 2014.

[33] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In Proceedings of the 30th International Conference on Machine Learning (ICML), Vol. 28, pp. 1310–1318, 2013.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In Proceedings of CVPR, pp. 2818–2826, 2016.

[35] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In Proceedings of the 3rd Conference on Machine Translation (WMT), pp. 1–9, 2018.

[36] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

A 付録

表 3 にテストセットに含まれる分野、対訳文数および英語側単語数を示す。また、表 4 に分野別学習データに含まれる文数および英語側単語数を示す。なお、ASPEC は本来約 300 万文の学習データを含んでいるが、先行研究に基づき先頭 200 万文のみを学習に使用した [30]。翻訳モデルを学習する際のハイパーパラメータを表 5 に示す。

表 3 テストセットに含まれる分野、対訳文数および英語側単語数

テストセット	分野	文数	単語数
ASPEC [16]	科学技術論文	1,812	39,573
JESC [17]	映画字幕	2,000	13,617
KFTT [18]	Wikipedia 記事	1,160	22,063
TED (tst2015) [19]	TED Talk	1,194	20,367
Business Scene Dialogue Corpus [20]	対話	2,120	19,619
WMT20 News En-Ja [21]	ニュース	1,000	22,141
WMT20 News Ja-En [21]	ニュース	993	24,423
WMT21 News En-Ja [22]	ニュース	1,000	23,305
WMT21 News Ja-En [22]	ニュース	1,005	24,771
WMT19 Robustness En-Ja (MTNT2019) [23]	Reddit	1,392	19,988
WMT19 Robustness Ja-En (MTNT2019) [23]	Reddit	1,111	13,390
WMT20 Robustness Set1 En-Ja [24]	Wikipedia コメント	1,100	29,419
WMT20 Robustness Set2 En-Ja [24]	Reddit	1,376	20,011
WMT20 Robustness Set2 Ja-En [24]	Reddit	997	15,866
IWSLT21 Simultaneous Translation En-Ja Dev [25]	TED Talk	1,442	20,677

表 4 分野別学習データに含まれる文数および英語側単語数。ASPEC は本来約 300 万文の学習データを含んでいるが、先行研究に基づき先頭 200 万文のみを学習に使用した [30]。

データ	文数	単語数
ASPEC	3,008,500	68,929,413
JESC	2,797,388	19,339,040
KFTT	440,288	9,737,715
TED	223,108	3,877,868

表 5 ハイパーパラメータの一覧

共通設定	
Architecture	Transformer [4]
Enc-Dec Layers	6
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$) [31]
Learning Rate Schedule	Inverse square root decay
Warmup Steps	4,000
Max Learning Rate	0.001
Dropout	0.3 [32]
Gradient Clipping	1.0 [33]
Label Smoothing	$\epsilon_{ls} = 0.1$ [34]
Mini-batch Size	512,000 tokens [35]
Number of Updates	36,000 steps (v3.0), 24,000 steps (v1,0, v2.0)
Averaging	100 ステップごとにモデルを保存し、最終 8 チェックポイントの平均を用いる
Beam Size	6 (文長による正規化付き) [36]
Small 設定	
Attention Heads	4
Word-embedding Dimension	512
Feed-forward dimension	1,024
Base 設定	
Attention heads	8
Word-embedding dimension	512
Feed-forward dimension	2,048
Big 設定	
Attention heads	16
Word-embedding dimension	1,024
Feed-forward dimension	4,096